SentBench: Comprehensive Evaluation of Self-Supervised Sentence Representation with Benchmark Construction

Xiaoming Liu[1,3], **Hongyu Lin**[1*], **Xianpei Han**[1,2*], **Le Sun**[1,2]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
{xiaoming2021,hongyu,xianpei,sunle}@iscas.ac.cn

Abstract

Self-supervised learning has been widely used to learn effective sentence representations. Previous evaluation of sentence representations mainly focuses on the limited combination of tasks and paradigms while failing to evaluate their effectiveness in a wider range of application scenarios. Such divergences prevent us from understanding the limitations of current sentence representations, as well as the connections between learning approaches and downstream applications. In this paper, we propose SentBench, a new comprehensive benchmark to evaluate sentence representations. SentBench covers 12 kinds of tasks and evaluates sentence representations with three types of different downstream application paradigms. Based on SentBench, we re-evaluate several frequently used self-supervised sentence representation learning approaches. Experiments show that SentBench can effectively evaluate sentence representations from multiple perspectives, and the performance on SentBench leads to some novel findings which enlighten future researches.

1 Introduction

Self-supervised representation learning is considered an important reason for breakthroughs in NLP (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019). And learning effective sentence representations has long been a fundamental challenge. (Kiros et al., 2015; Conneau et al., 2017; Cer et al., 2018). In recent years, various self-supervised sentence representation learning approaches leverage different self-constrained signals, e.g., sentence pairs in the same narratives (Devlin et al., 2019), sentence order (Lan et al., 2019), or sentence permutation (Lewis et al., 2020), to learn representations by training models to distinguish positive instances from negatives.

Even though current self-supervised sentence representation approaches have reached significant progress on some datasets like Semantic Textual Similarity (STS) (Ho and Nvasconcelos, 2020; Gao et al., 2021), benchmarks for evaluation lag far behind the development of methods (Wang et al., 2022). Currently, sentence representations are evaluated in limited tasks and specific paradigms. For example, the most commonly used SentEval benchmark (Conneau and Kiela, 2018) mainly focuses on single sentence classification and semantic similarity tasks. Unfortunately, prior literature shows that performance on STS cannot reflect the effectiveness of sentence representations on a wider range of tasks (Reimers et al., 2016; Zhelezniak et al., 2019; Wang et al., 2022). And available evaluation toolkits assess the same downstream task with a singular paradigm, limiting our perception of methods in different application scenarios. Moreover, current self-supervised sentence representation learning approaches are coupled with multiple factors, including diverse contrastive signals, training losses, and model architectures. Consequently, evaluating whether, where, and how a learning method will benefit the downstream tasks is difficult.

In this paper, we propose SentBench, a new benchmark to comprehensively evaluate sentence representations with various downstream tasks and evaluation paradigms. As shown in Figure 1, SentBench

*Corresponding authors.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

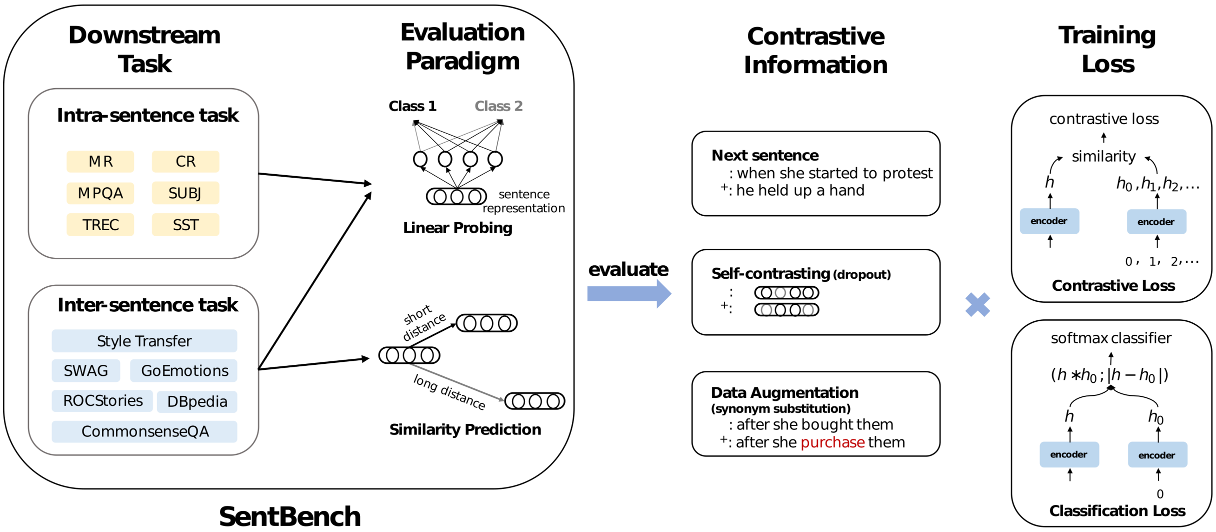

Figure 1: The framework of the paper (SentBench and decoupling analysis scheme).

contains 12 kinds of NLP tasks, including sentiment classification, question answering, story cloze, etc., and three evaluation paradigms, including single sentence classification, sentence pair classification and sentence pair contrasting (Zhu et al., 2018). The classification paradigm trains a simple additional classifier to assess information within representations for single sentence tasks or identify the connection between two candidate representations for pair-wise tasks. Besides, contrasting paradigm is similar to common retrieval or ranking scenario. Finally, SentBench constructs 18 datasets, which cover diverse tasks and common applications of sentence representations.

Based on SentBench, we re-evaluate several widely used self-supervised sentence representation learning approaches. We decouple previous approaches from two perspectives to identify critical factors: contrasting knowledge applied to construct positive instances and training losses used to optimize models. Specifically, we concentrate on three contrasting knowledge, including next sentence prediction (Devlin et al., 2019), self-contrasting (Yan et al., 2021; Gao et al., 2021) and data augmentation (Zhang et al., 2015; Feng et al., 2021), as well as two widespread training losses, including contrastive loss and classification loss. By thoroughly comparing different approaches on SentBench, we find that the advantages of the state-of-the-art methods can not be exhibited consistently to a broader range of downstream tasks and evaluation paradigms. Furthermore, the applied training loss leads to more significant impacts than contrasting knowledge. These findings shed some light on future research on sentence representation learning.

2 Benchmark Construction

2.1 Tasks

SentBench covers 12 downstream tasks for evaluating sentence representations, divided into single sentence and sentence pair tasks. In the following, we will briefly describe tasks in SentBench.

Single sentence tasks aim to classify sentence representations into corresponding categories. Because the previous SentEval[0] benchmark has covered extensive single sentence classification tasks, SentBench inherits all of them, including sentiment analysis (MR, SST) (Pang and Lee, 2005; Socher et al., 2013), Opinion Polarity (MPQA, SUBJ) (Wiebe et al., 2005; Pang and Lee, 2004), Question type (TREC) (Voorhees and Tice, 2000), product reviews (CR) (Hu and Liu, 2004).

Sentence pair tasks aim to identify sentence pairs with specific connections. We investigate six tasks covering various fields of downstream applications of NLP (Table 1):

[0] https://github.com/facebookresearch/SentEval

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Dataset	Classification			Contrasting
	Train size	Valid Size	Test Size	
SWAG	56,131	18,711	18,711	20,006
DBpedia	89,965	27,988	27,989	69,971
GoEmotions	54,535	18,178	18,179	4,590
ROCStories	2,513	-	629	1,571
StyleTransfer	24,986	8,328	8,330	2,500
CommonsenseQA	13,154	4,384	4,386	1,221

Table 1: The statistics of sentence pair tasks.

- **DBpedia** (Zhang et al., 2015), which identifies whether a pair of sentences come from the same category;
- **Style Transfer** (ST) (Jhamtani et al., 2017), which distinguishes whether modern English and Shakespearean English expresses same content;
- **GoEmotions** (GoEmo) (Demszky et al., 2020), which recognizes whether a sentence pair expresses similar fine-grained emotion;
- **ROCStories** (ROC) (Mostafazadeh et al., 2016), which predicts whether a given sentence is the proper ending to a four-sentence story;
- **CommonsenseQA** (CQA) (Talmor et al., 2019), which determines if candidate answers match a commonsense question;
- **SWAG** (Zellers et al., 2018), which predicts correct answer for a question about grounded situations.

2.2 Evaluation paradigm

We design three evaluation paradigms in SentBench:

- **single sentence classification** directly leverage sentence representations as features with a simple classifier to assess how much desirable information is contained in representations;
- **sentence pair classification** trains a simple classifier that determines whether there is a specific connection between candidate sentences, that is mapping a pair of sentence representation $(\mathbf{x_1}, \mathbf{x_2})$ into corresponding label;
- **sentence pair contrasting** distinguishes a sentence from candidates that are more likely to share a specific relationship with the given sentence, i.e., given a target sentence x and two candidates (x^+, x^-), sentence pair contrasting selects more suitable candidate based on the similarity between x, x^+, and x^-.

Note that the classification paradigm requires data to train additional classifier parameters, while sentence pair contrasting depends on the similarity between sentence pairs by directly calculating certain distance metrics (e.g., cosine similarity) without additional training instances. Therefore, we provide training and development sets for classification tasks.

3 Experiment Setup

Based on SentBench, we re-evaluate several most frequently used self-supervised sentence representation methods. Since contrasting knowledge and training losses are usually coupled, it is challenging to directly identify critical factors for successful sentence representations from previous works. To this end, this paper explores different combinations of contrasting knowledge and training losses to investigate the effects of distinct factors.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Contrasting Knowledge. We exploit three popular contrasting knowledge sources:

- **narrative contrasting**, which predicts whether a hypothesis sentence belongs to the same narrative with a premise, is also known as next sentence prediction (NSP);
- **self-contrasting**, which disturbs sentence representations at feature-level, tries to distinguish representations stemming from the same instance. SimCSE (Gao et al., 2021) is one of the most popular methods, which creates contrasting pairs via random dropout from neural networks;
- **data augmentation**, which modifies the original instances via some rule-based modification, and tries to distinguish original instances from others.

In this paper, we apply NSP (Devlin et al., 2019), two-times Dropout (Dropout) (Gao et al., 2021), and synonym substitution (DA) (Wu et al., 2020) as each knowledge sources, respectively.

Training Loss. Contrastive loss and classification loss are the most popular loss functions in self-supervised sentence representation learning. Given an instance $\mathbf{x}$, **contrastive loss** (CTR) (Van den Oord et al., 2018) aims to distinguish positive instance representation $\mathbf{x}^+$ from a batch of negatives:

$$\mathcal{L}_{CTR}(\theta) = -\log \frac{e^{sim(\mathbf{x},\mathbf{x}^+)/\tau}}{\sum_{\mathbf{x}_i \in batch} e^{sim(\mathbf{x},\mathbf{x}_i)/\tau}}$$

where τ is a temperature hyperparameter and sim is a similarity function (e.g., cosine similarity).

classification loss (CLS) classifies sentence pairs representation into corresponding semantic labels:

$$\begin{aligned}\mathcal{L}_{CLS}(\theta) = &-\log P(y = 1 \big| \mathbf{x} * \mathbf{x}^+) \\ &- \sum_{\mathbf{x}^- \in batch} \log P(y = 0 \big| \mathbf{x} * \mathbf{x}^-)\end{aligned}$$

where $*$ is the concatenation of representations.

Implementation Details. We implement the above-mentioned approaches based on BERT$_{\texttt{base}}$ (uncased) (Lan et al., 2019) and RoBERTa$_{\texttt{base}}$ (Liu et al., 2019). To compare the benefit of different approaches, we also implement two token-aggregation approaches without further learning as baselines, which regard average representations of all tokens or the [CLS][1] representation of the last layer of models as sentence representation.

In this paper, we use BookCorpus (Zhu et al., 2015) to construct the next sentence samples. Devlin et al. (2019) concatenate two sentences with [SEP] and feed the [CLS] representation into the classifier. A slight difference from the above approach is that we first obtain the [CLS] representations of two sentences separately and then concatenate them to learn the next sentence prediction. For self-supervised sentence representation learning with different combinations of loss functions and contrasting knowledge, we train models for one epoch on 10^6 sentences from BookCorpus and set batch size to 64. The temperature τ of contrastive loss is set to 0.05, and max sequence length is set to 32. Cosine similarity is the default distance metric and similarity function. All experiments are run in NVIDIA TITAN RTX GPUs. Following Gao et al. (2021) and Wu et al. (2020), the best checkpoint on the development set of STS is saved for evaluation. We use NLPAUG[2] for synonym substitution and take other sentences in the same mini-batch as negatives.

4 Empirical Findings

Table 2, 3 and 4 show the experiment results on three evaluation paradigms in SentBench, respectively. From these empirical results, we obtain the following findings.

[1]We discard the MLP layer over [CLS] for evaluation.
[2]`https://github.com/makcedward/nlpaug`

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Model	MR	CR	MPQA	SUBJ	SST	TREC	AVG
BERT-AVG	**82.24**^{1}	87.39^{1}	**88.71**^{2}	95.45^{3}	84.62^{4}	**91.80**^{1}	88.37^{2}
BERT-[CLS]	81.83^{2}	**87.39**^{1}	88.21^{6}	**95.48**^{2}	**86.91**^{1}	91.33^{2}	**88.53**^{1}
Dropout (CTR)	80.43^{4}	85.09^{5}	88.43^{4}	94.64^{6}	84.66^{3}	**90.67**^{3}	87.32^{4}
Dropout (CLS)	67.73^{8}	70.09^{8}	85.50^{7}	87.93^{8}	75.36^{8}	79.33^{8}	77.66^{8}
NSP (CTR)	**81.13**^{3}	**87.18**^{3}	88.34^{5}	**95.53**^{1}	**85.05**^{2}	89.67^{5}	**87.82**^{3}
NSP (CLS)	78.92^{6}	85.59^{4}	88.54^{3}	95.10^{4}	83.42^{6}	89.87^{4}	86.91^{6}
DA (CTR)	80.16^{5}	84.64^{6}	**89.33**^{1}	94.72^{5}	83.98^{5}	89.67^{5}	87.08^{5}
DA (CLS)	73.89^{7}	77.25^{7}	80.10^{8}	90.74^{7}	77.46^{7}	84.73^{7}	80.70^{7}
RoBERTa-AVG	**83.43**^{3}	**88.58**^{2}	**86.75**^{5}	**95.22**^{2}	**87.26**^{3}	**91.93**^{1}	**88.80**^{2}
RoBERTa-[CLS]	81.27^{4}	86.01^{5}	84.18^{6}	94.15^{4}	86.66^{4}	83.00^{6}	85.88^{6}
Dropout (CTR)	80.18^{5}	85.43^{6}	87.55^{2}	93.22^{6}	85.35^{5}	87.80^{5}	86.59^{5}
Dropout (CLS)	60.58^{7}	63.84^{8}	77.82^{7}	81.10^{7}	70.45^{7}	66.60^{7}	70.07^{7}
NSP (CTR)	**85.90**^{1}	**90.60**^{1}	**88.96**^{1}	**95.39**^{1}	**91.12**^{1}	**91.33**^{2}	**90.55**^{1}
NSP (CLS)	83.62^{2}	88.51^{3}	87.51^{3}	94.72^{3}	87.75^{2}	89.67^{3}	88.63^{3}
DA (CTR)	80.03^{6}	86.78^{4}	87.12^{4}	93.23^{5}	84.47^{6}	89.13^{4}	86.79^{4}
DA (CLS)	56.02^{8}	63.97^{7}	74.10^{8}	77.59^{8}	61.25^{8}	65.60^{8}	66.42^{8}

Table 2: Accuracies on single sentence classification tasks and corner markers represent the performance rank. CTR: contrastive loss; CLS: classification loss.

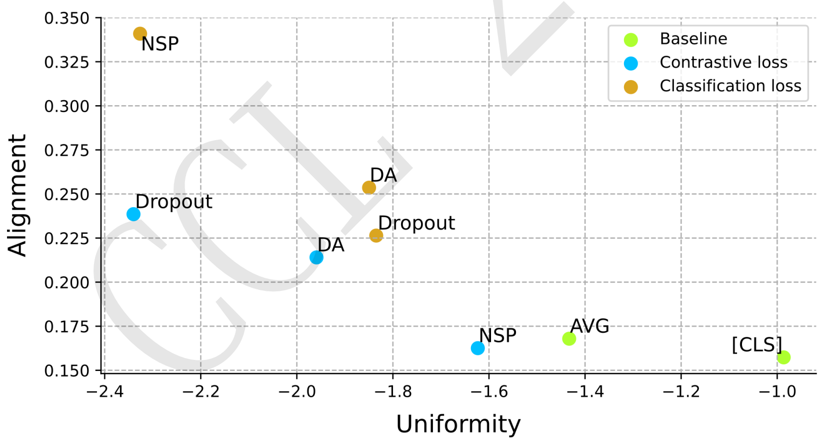

Figure 2: Alignment and uniformity plot of models based on BERT. For both alignment and uniformity, lower numbers are better.

Finding 1. Training loss is a more critical factor than contrasting knowledge. We find that the selection of training loss has more significant impacts than the selection of contrasting knowledge, and contrastive loss significantly outperforms classification loss across all contrasting knowledge, models, and evaluation paradigms. Note that previously NSP is commonly coupled with classification loss and therefore achieves little performance superiority (Liu et al., 2019). However, from our experiments, NSP trained with contrastive loss can bring significant performance improvements. To further investigate how contrasting knowledge and training loss influence sentence representations, we calculate the alignment and uniformity, two quantified quality evaluation metrics for sentence representations

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

(Wang and Isola, 2020). As shown in Figure 2, we can see that different contrasting information is essentially a trade-off between alignment and uniformity. And contrastive loss outperforms classification loss with better alignment and uniformity, which reveals the underlying reason for the superior performances.

Model	ST	DBpedia	GoEmo	ROC	CQA	SWAG	AVG
BERT-AVG	**86.03**^{1}	91.35^{6}	**56.64**^{5}	**63.12**^{2}	**58.38**^{3}	**65.81**^{2}	**70.22**^{2}
BERT-[CLS]	85.76^{3}	**91.57**^{5}	56.51^{6}	60.15^{4}	54.30^{6}	64.19^{3}	68.75^{6}
Dropout (CTR)	84.19^{6}	92.29^{4}	57.33^{3}	56.60^{6}	59.69^{2}	62.52^{5}	68.77^{5}
Dropout (CLS)	79.19^{8}	79.83^{8}	52.18^{8}	53.58^{8}	50.97^{7}	52.94^{8}	61.45^{8}
NSP (CTR)	**85.93**^{2}	**96.07**^{1}	**59.06**^{1}	**64.07**^{1}	**60.11**^{1}	**66.05**^{1}	**71.88**^{1}
NSP (CLS)	84.40^{5}	95.67^{2}	57.18^{4}	59.41^{5}	55.88^{5}	63.41^{4}	69.34^{4}
DA (CTR)	84.92^{4}	93.34^{3}	57.78^{2}	61.05^{3}	57.83^{4}	61.08^{6}	69.33^{3}
DA (CLS)	80.42^{7}	83.60^{7}	53.06^{7}	54.00^{7}	50.82^{8}	54.83^{7}	62.79^{7}
RoBERTa-AVG	**83.41**^{3}	89.17^{6}	**54.90**^{5}	**59.46**^{4}	**54.43**^{5}	**65.91**^{1}	**67.88**^{4}
RoBERTa-[CLS]	81.60^{5}	**89.78**^{5}	53.76^{6}	55.12^{6}	50.47^{7}	64.22^{2}	65.83^{6}
Dropoput (CTR)	82.09^{4}	92.74^{4}	55.38^{4}	55.75^{5}	56.72^{2}	60.46^{5}	67.19^{5}
Dropout (CLS)	75.16^{7}	69.62^{7}	50.72^{7}	53.15^{8}	49.93^{8}	51.76^{7}	58.39^{7}
NSP (CTR)	**84.83**^{1}	**96.49**^{1}	**58.95**^{1}	**66.93**^{1}	**60.41**^{1}	**63.85**^{3}	**71.91**^{1}
NSP (CLS)	83.46^{2}	95.74^{2}	56.70^{3}	63.01^{2}	55.82^{4}	61.54^{4}	69.38^{2}
DA (CTR)	81.47^{6}	94.69^{3}	57.88^{2}	59.62^{3}	55.83^{3}	59.15^{6}	68.11^{3}
DA (CLS)	74.12^{8}	66.97^{8}	50.16^{8}	53.21^{7}	50.52^{6}	51.30^{8}	57.71^{8}

Table 3: Accuracies on sentence pair classification tasks and corner markers represent the performance rank. CTR: contrastive loss; CLS: classification loss.

Finding 2. Narrative contrasting provides more useful information for a wide range of single sentence and sentence pair tasks. Experiments show that the NSP with contrastive loss achieves satisfactory performance in almost all settings. Besides, we can see that performance improvement on RoBERTa is more significant than that of BERT. This may be because the [CLS] representation of BERT has been pretrained with NSP signals and therefore already contain such kind of knowledge. Furthermore, we find that self-contrasting strategies, which are reported to achieve superior performance on STS benchmarks (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015; Agirre et al., 2016), do not perform well in SentBench. We believe that this is because, as previous findings have shown (Wang et al., 2022), STS tasks have a weak correlation with downstream tasks. Therefore, evaluations on STS benchmarks are not universal, revealing the necessity of building SentBench.

Finding 3. Self-supervised contrastive sentence representation learning leads to more significant improvements on sentence pair contrasting tasks. We can see that for BERT-AVG and RoBERTa-AVG, there are 6.2% and 12% of average performance improvements of all methods with contrastive loss, which is significantly higher than that on the other two tasks. We speculate that contrastive loss is more appropriate for similarity-based evaluation, which substantially improves the consistency between sentence representation distribution and downstream applications. Furthermore, single sentence and sentence pair classification tasks introduce an additional trainable classifier, which may weaken the effectiveness of self-supervised pretraining. Consequently, self-supervised contrastive sentence representation is more suitable for similarity-based scenarios without additional supervised signals, which is also consistent with recent advances of these methods on previous STS benchmarks (Gao et al., 2021).

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Model	ST	DBpedia	GoEmo	ROC	CQA	SWAG	AVG
BERT-AVG	63.88^{8}	**85.89**^{5}	**57.02**^{4}	58.75^{4}	**52.99**^{5}	**56.50**^{5}	**62.50**^{5}
BERT-[CLS]	**65.52**^{6}	74.72^{6}	53.09^{6}	**59.90**^{3}	52.09^{6}	54.19^{6}	59.92^{6}
Dropout (CTR)	73.16^{2}	91.43^{4}	57.56^{2}	60.53^{2}	**67.49**^{1}	62.01^{2}	68.70^{2}
Dropout (CLS)	73.16^{2}	66.53^{7}	52.96^{7}	52.45^{8}	51.68^{7}	51.30^{7}	58.01^{7}
NSP (CTR)	71.84^{4}	**94.68**^{1}	**57.82**^{1}	**62.70**^{1}	65.85^{3}	**63.24**^{1}	**69.35**^{1}
NSP (CLS)	64.72^{7}	94.62^{2}	56.27^{5}	56.02^{6}	61.51^{4}	57.48^{4}	65.10^{4}
DA (CTR)	**75.52**^{1}	91.76^{3}	57.47^{3}	57.73^{5}	66.83^{2}	59.64^{3}	68.16^{3}
DA (CLS)	71.48^{5}	64.02^{8}	52.14^{8}	52.51^{7}	49.80^{8}	50.86^{8}	56.80^{8}
RoBERTa-AVG	61.20^{8}	67.91^{6}	50.11^{8}	52.13^{8}	55.61^{6}	51.32^{6}	56.38^{7}
RoBERTa-[CLS]	**73.96**^{3}	**86.20**^{5}	**51.90**^{5}	**58.82**^{5}	**56.35**^{5}	**60.32**^{3}	**64.59**^{5}
Dropout (CTR)	**75.68**^{1}	90.76^{4}	55.88^{3}	60.09^{3}	64.86^{2}	61.98^{2}	68.21^{2}
Dropout (CLS)	70.60^{6}	63.38^{7}	51.35^{6}	56.72^{6}	52.25^{7}	49.94^{7}	57.37^{6}
NSP (CTR)	69.64^{7}	**96.78**^{1}	**58.26**^{1}	**64.74**^{1}	**65.44**^{1}	**62.96**^{1}	**69.64**^{1}
NSP (CLS)	70.80^{4}	95.17^{2}	55.53^{4}	63.91^{2}	61.43^{4}	59.77^{4}	67.77^{3}
DA (CTR)	74.76^{2}	94.17^{3}	57.71^{2}	59.01^{4}	61.92^{3}	57.00^{5}	67.43^{4}
DA (CLS)	70.72^{5}	59.48^{8}	50.13^{7}	52.32^{7}	46.85^{8}	49.75^{8}	54.87^{8}

Table 4: Accuracies on sentence pair contrasting tasks and corner markers represent the performance rank.

5 Related Works

SentEval vs SentBench SentEval and SentBench are both benchmarks that evaluate the quality of sentence representations in natural language processing tasks. SentEval consists of a set of 17 downstream tasks and 10 probe tasks, including sentiment analysis, natural language inference, paraphrase detection, and text similarity. However, the tasks and methods in SentEval have fallen behind in recent years due to the rapid development of models and methods.

SentBench builds on SentEval, expanding the sentence-pair tasks to include six new datasets such as commonsense QA, story generation, and fine-grained sentiment analysis. Previous studies have shown that the performance of text semantic similarity tasks cannot reflect the effectiveness of sentence representations in more downstream tasks (Reimers et al., 2016; Zhelezniak et al., 2019; Wang et al., 2022). Unlike SentEval, SentBench replaces text similarity tasks with contrasting tasks, which can more objectively reflect the actual application performance of sentence representations. Additionally, SentBench adds different evaluation paradigms to enrich the evaluation forms of the data, which can provide different understanding perspectives for the same downstream task.

GLUE vs SentBench The General Language Understanding Evaluation (GLUE) benchmark is a collection of nine natural language processing tasks designed to assess the performance of language models in various natural language understanding tasks, including sentiment analysis, question answering, and natural language inference. Unlike SentBench, which aims to evaluate sentence representation models and methods, GLUE is designed to evaluate and analyze natural language understanding systems. Although both benchmarks contain sentence representation-related applications, the differences in their design goals result in differences in datasets and usage methods. While SentBench focuses on the generalization and universality of sentence representations, GLUE tests the overall ability of the model. Additionally, the datasets used in GLUE and SentBench are complementary, as SentBench does not currently collect data relevant to natural language inference tasks. Thus, SentBench could look to GLUE's

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

relevant content for future expansion.

Probing Researchers have not only focused on building more efficient evaluation benchmarks but also used various probing tasks to uncover the underlying principles of sentence representation, such as identifying syntactic and semantic information, as well as subtle perturbations. These evaluation tasks offer insights into which factors are challenging for sentence representation and which can better distinguish different models, driving the development of sentence representation. In their attempt to analyze sentence representation, Adi et al. (2016) designed three evaluation tasks that focused only on surface information, such as sentence length, sentence content, and word order, and experimented with popular methods. However, these evaluation tasks failed to reflect the syntactic, semantic, and other knowledge of sentence representation. To address this limitation, Conneau et al. (2018) designed and collected 10 probing tasks that were divided into categories of surface, syntactic, and semantic information, revealing differences and connections between different methods. Furthermore, Zhu et al. (2018) proposed a triplet evaluation framework that generated triplet sentences to explore how syntactic structure or semantic changes in a given sentence affected inter-sentence similarity. This approach not only evaluated the performance of different sentence representation methods in capturing different semantic attributes but also avoided bias from human annotation data, providing a better understanding of these methods. Our work is similar to the previously mentioned research in that we aim to investigate the underlying mechanisms of sentence representation learning through thorough more comprehensive evaluation and decoupling analysis.

6 Conclusion

In this paper, we propose a new universal sentence evaluation benchmark SentBench, which introduces more downstream tasks and evaluation paradigms. Furthermore, we decouple and analyze the effects of contrasting knowledge and training losses on sentence representations. Empirical findings show that training losses play a more critical role in self-supervised sentence representation learning and help us better understand and design sentence representation learning algorithms.

7 Limitations

Currently, SentBench mainly covers English datasets, and therefore can not evaluate whether self-supervised representation learning methods have some language-specific properties. Besides, due to the limitation of time, we mainly experiment with BERT and RoBERTa without evaluating more self-supervised sentence representations methods, such as Sentence-T5 (Ni et al., 2022). Finally, we mainly focus on the performance of models on SentBench without discussing more details of the training process, which is also an important aspect of self-supervised sentence representations.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. U1936207, 62122077 and 62106251.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM),* *Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, August. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California, June. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, August. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Chih-Hui Ho and Nuno Nvasconcelos. 2020. Contrastive learning with adversarial examples. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17081–17093. Curran Associates, Inc.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland, May. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain, July.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.

Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.

Ellen M Voorhees and Dawn M Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

Bin Wang, C.-C. Jay Kuo, and Haizhou Li. 2022. Just rank: Rethinking evaluation with word and sentence similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077, Dublin, Ireland, May. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online, August. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium, October-November. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. Correlation coefficients and semantic textual similarity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 951–962, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December.

Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637, Melbourne, Australia, July. Association for Computational Linguistics.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 813-823, Harbin, China, August 3 - 5, 2023.